Journal of the Faculty of Engineering and Architecture of Gazi University 39:2 (2024) 709-716

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Analysis of Turkish audio recording data labeled with three emotions popular machine learning algorithms

Abdulkadir Tepecik[1]*, Engin Demir[2]
[1]Computer Engineering Department, Engineering Faculty, Yalova University, 77200, Yalova, Türkiye
[2]Department of Computer Technology, National Defense University, Balıkesir, Türkiye

Highlights:

- Turkish emotion data set has been created.
- The Bert model was used on Turkish Data
- Emotion analysis of Turkish voice recording data was performed not directly through voice, but through text states.

Keywords:

- Sentiment Analysis
- Turkish Audio Recordings
- Machine Learning
- Turkish Sentiment Analysis
- Rapid Miner
- Python

Article Info:

Research Article
Received: 12.09.2021
Accepted: 12.04.2023

DOI:

10.17341/gazimmfd.994478

Correspondence:

Author: Abdulkadir Tepecik
e-mail: atepecik@yalova.edu.tr
phone: +90 226 815 5453

Graphical/Tabular Abstract

Emotions are physical changes in a person's mood resulting from his interaction with internal and environmental influences. Individuals can convey their feelings to other individuals by means of voice communication as well as body language. Voice communication becomes important for individuals, especially in a situations and times when body language is insufficient. In our study, first of all, three emotion tags were determined on the data set containing the Python programming language and Turkish voice recordings, and then analyses were carried out with the five most used machine learning algorithms in the literature studies (Figure A).

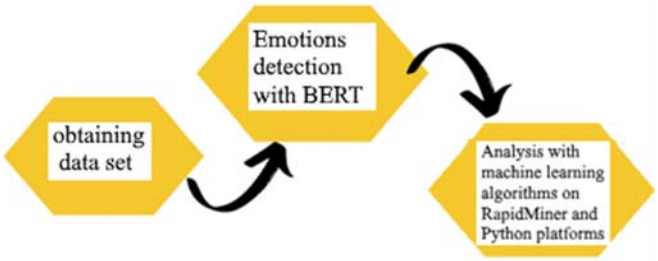


Figure A. Flow Chart of the Study

Purpose:
With our study, it is aimed to create a Turkish emotion database, to lead the studies to be carried out with methods and materials, and to develop the Turkish structure of the BERT model used in emotion detection. It is aimed to measure the success rates of machine learning algorithms on Turkish language structure.

Theory and Methods:
Sentiment detection and analysis of Turkish voice recordings are aimed. For this study; the BERT model and Python programming language were used in the emotion detection stage. For the analysis part of the study, the five most used machine learning algorithms in the literature review were determined. These algorithms are Naive Bayes, Support Vector Machine, Random Forest, Decision Tree, and K-Nearest Neighborhood. As a vectorization method, both CountVectorizer and TF-IDF methods were used in Python, and TF-IDF method was used in RapidMiner. Python programming language and Rapid Miner were used for analysis using machine learning algorithms.

Results:
As a result, Naive Bayes and Support Vector Machine algorithms obtained the best accuracy rate in the Python programming language with 67%. In RapidMiner, the Naïve Bayes machine learning algorithm achieved the best accuracy rate of 60.61%.

Conclusion:
Advances in technology have allowed studies in data and data science to be carried out in more advanced environments and facilities. Especially data analysis has increased in popularity recently. In our study, analyses were made with emotion detection and machine learning algorithms on the data set containing the data of Turkish voice recordings. The additive nature of the Turkish language structure and the fact that the data in the dataset contains mostly news texts have been effective in obtaining these results. Google's CoLab environment, which was used in the analysis with the Python programming language, offered wider possibilities compared to the RapidMiner platform. Faster results were obtained compared to RapidMiner and enabled the use of more than one method. This advantage provided a better analysis of the study and contributed to more appropriate evaluations of the results.

Journal of the Faculty of Engineering and Architecture of Gazi University 39:2 (2024) 709-716

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Üç duygu ile etiketlenmiş Türkçe ses kayıt verilerinin makine öğrenim algoritmalarıyla analizi

Abdulkadir Tepecik[1]*, Engin Demir[2]
[1]Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 77200, Yalova, Türkiye
[2]Milli Savunma Üniversitesi, Bilgisayar Teknolojileri Bölümü, Balıkesir, Türkiye

ÖNEÇIKANLAR

- Türkçe duygu veri seti oluşturulmuştur
- BERT modeli Türkçe veriler üzerinde kullanılmıştır
- Türkçe ses kayıt verilerinin direkt ses üzerinden değil, metin halleri üzerinden duygu analizi yapılmıştır

Makale Bilgileri
Araştırma Makalesi
Geliş: 12.09.2021
Kabul: 12.04.2023

DOI:
10.17341/gazimmfd.994478

Anahtar Kelimeler:
Duygu Analizi,
makine öğrenmesi,
Türkçe ses kayıtları,
Türkçe duygu analizi,
rapidminer,
python

ÖZ

Teknolojinin gelişimiyle beraber gündelik hayatımızdaki birçok işlem internet ortamında gerçekleşmektedir. Her geçen gün artan internet kullanımı da bu ortamdaki veri sayısını artırmıştır. Verilerin işlenmesi, analizi gibi konular özellikle bilişim dünyasında yeni bilim alanlarının oluşmasına imkân sağlamıştır. Çalışmamızda da Türkçe ses kayıtlarını içeren veri seti üzerinde Python programlama dili aracılığıyla öncelikle verilerin duygu etiketlerinin tespiti yapılmış olup, sonrasında literatür çalışmalarında en çok kullanılan beş makine öğrenim algoritmasıyla analizler gerçekleştirilmiştir. Analizler hem RapidMiner hem de Python programlama dili aracılığıyla gerçekleştirilmiştir. Çalışmada, Python programlama dili aracılığıyla yapılan analizlerde hem CountVectorizer hem de TF-IDF vektörizasyon yöntemleri, RapidMiner ile yapılan analizlerde TF-IDF vektörizasyon yöntemi kullanılmıştır. Sonuç kısmında ise Python programlama dilinde en iyi doğruluk oranını %67 ile Naive Bayes ve Destek Vektör Makinesi algoritmaları elde etmiştir. RapidMiner'da ise en iyi doğruluk oranını %60,61 oranla Naïve Bayes makine öğrenim algoritması elde etmiştir. Çalışmamızla beraber ortaya yeni bir Türkçe duygu veri seti çıkmıştır. Çalışmamız ayrıca Türkçe ses kayıtlarından elde edilen verilerin BERT modeli ile duygu tespiti yapılan özgün bir çalışmadır.

Analysis of Turkish audio recording data labeled with three emotions popular machine learning algorithms

HIGHLIGHTS

- Turkish emotion data set has been created
- The Bert model was used on Turkish Data
- Emotion analysis of Turkish voice recording data was performed not directly through voice, but through text states

Article Info
Research Article
Received: 12.09.2021
Accepted: 12.04.2023

DOI:
10.17341/gazimmfd.994478

Keywords:
Sentiment analysis,
Turkish sentiment analysis,
machine learning,
Turkish auido recordings,
RapidMiner,
python

ABSTRACT

With the development of technology, many transactions in our daily life take place on the internet. The increasing use of the internet has also increased the number of data in this environment. Subjects such as processing and analysis of data have allowed the formation of new fields of science, especially in the world of informatics. In our study, the emotion labels of the data were firstly determined by the Python programming language on the data set containing the Turkish voice recordings, and then the analyzes were carried out with the five most used machine learning algorithms in the literature studies. Analyses were conducted through both Rapid Miner and the Python programming language. In the study, both CountVectorizer and TF-IDF vectorization methods were used in analyses performed through the Python programming language, and TF-IDF vectorization method was used in analyses performed with Rapid Miner. As a result, Naive Bayes and Support Vector Machine algorithms obtained the best accuracy rate in the Python programming language with 67%. In RapidMiner, the Naïve Bayes machine learning algorithm achieved the best accuracy rate of 60.61%. Our study is also an original study in which emotion detection was done with the BERT model of the data obtained from Turkish voice recordings.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *atepecik@yalova.edu.tr, engindmr7@gmail.com / Tel: +90 226 815 5453

1. Giriş (Introduction)

Kitle iletişim araçlarının gelişimi hayatımıza yeni bir boyut kazandırmıştır. Günümüzde internet ve internetin oluşturduğu sanal ortam, bireyler için artık vazgeçilmez hale gelmiştir. Kişiler duygularını, düşüncelerini artık bu ortam sayesinde aktarabilmekte ve bunları paylaşabilmektedir. Bu paylaşımlar, birden fazla verinin oluşmasına da katkı sağlamaktadır.

Veri ve veri bilimi, son zamanların popüler araştırma alanları haline getirmiştir. Sanal ortamda oluşan veri yığınları, çeşitli çalışmalar için eşsiz bir kaynak haline gelmiştir. Bu durum yeni çalışma alanlarını ortaya çıkardığı gibi yeni bilim alanlarını da ortaya çıkarmıştır.

Veri biliminin ve veri analizinin en etkin kullanıldığı alanlardan biri olan makine öğrenmesi, *bir taraftan insan zekâsını ve algısını taklit ederken, diğer taraftan da kişinin yorumlayıp elle gireceği kurallara ihtiyaç duymayan algoritmalar bütünü* olarak ifade edilmektedir [1]. Makine öğrenimi içerdiği algoritmalar sayesinde birçok metrik üzerinde değerlendirilme yapılmasına olanak sağlamıştır.

Türkçe ses kayıt verilerinin metinleri üzerinde yaptığımız çalışmamızın amacı; Türkçe duygu veri seti oluşturmak, BERT modelinin Türkçe veriler üzerindeki analiz başarısını tespit etmek ve bu model yapısıyla yapılacak olan çalışmalara öncü olmak ve makine öğrenim algoritmalarının Türkçe verilerdeki başarı oranlarını tespit etmek amaçlanmıştır.

Çalışmamızda öncelikle Türkçe ses kayıtlarını içeren veri setinin metin dokümanları elde edilmiştir. Elde edilen veri seti üzerinde anlamsız verilerin ayrıştırılması sonrasında 5001 anlamlı veri, duygu tespiti ve analiz işlemlerine hazır hale getirilmiştir. Veri seti üzerinde, verilerin duygu etiketleri BERT modeli aracılığıyla negatif, pozitif ve nötr olarak belirlenmiştir. Oluşan duygu veri tabanının analizi için literatür taramasında en çok kullanılan beş makine öğrenim algoritması belirlenmiştir. Bu algoritmalar; Naive Bayes, Destek Vektör Makinesi (*S*upport *V*ector *M*achine), Rastgele Orman(*R*andom *F*orest), Karar Ağacı(*D*ecision *T*ree) ve K-en Yakın Komşu(*K-N*earest *N*eighbour)'dur. Python programlama dili aracılığıyla gerçekleştirilen çalışmada yapılan analizlerin daha geniş değerlendirilmesi için hem CountVectorizer hem de TF-IDF vektörizasyon yöntemleri kullanılmıştır. Her iki vektörizasyon yöntemi, çalışmadaki makine öğrenim algoritmalarının hepsinde kullanılmıştır. RapidMiner platformuyla yapılan analizler ise TF-IDF vektörizasyon yöntemi ile değerlendirilmiştir. Sonuçlar ise doğruluk (accuracy), duyarlılık değeri (precision), anma değeri (recall) ve F1-Skor (F1-score) gibi metriklerle değerlendirilmiştir. Çalışmanın analizi hem RapidMiner platformunda hem de Python programlama dili ile Google'ın CoLab ortamında gerçekleştirilmiştir.

2. Literatür Çalışmaları (Literature Review)

İnsanların duygularını işitsel bilgilerden elde etmek için A. Demir vd. [2], otomatik duygu tanıma sistemi önermişlerdir. Çalışmada SAVEE, RAVDESS ve RML veri setleri kullanılarak bu veri setlerinde bulunan ses sinyallerinden özellikler çıkarılmış ve sınıflandırma için derin öğrenme tabanlı LSTM algoritması kullanılmıştır. Sonuç kısmında, önerilen yöntemle RAVDESS veri seti üzerinde %68,8, SAVEE veri seti üzerinde %72,13 ve RML veri seti üzerinde %70,35 doğruluk oranı elde edilmiştir. En başarılı sonuçların SAVEE veri seti için elde edildiğini ve bu durumun SAVEE veri setinin, ana dili İngilizce olan erkek konuşmacılardan elde edilmiş olmasından kaynaklandığını ifade etmişlerdir. İnsan konuşmalarındaki sesleri kullanarak duygu tanımlama sistemi için yeni bir yaklaşım sunan M. Kudiri vd. [3], çalışmada dört duygu (kızgın, üzgün, mutlu ve nötr) türünün tahmini sağlamaya çalışmışlardır. Bu çalışma içinde Berlin Duygu veri tabanını

kullanmışlardır. Sınıflandırma için Sinir Ağları ve Destek Vektör Makineleri makine öğrenim modelleri kullanılan çalışmanın sonuç kısmında Sinir Ağlarının ortalama %81, Destek Vektör Makinesinin ortalama %86 doğruluk oranı elde ettiği belirtilmiştir. Göreceli bin frekans katsayılı yaklaşımın umut verici sonuçlar verdiğini ifade etmişlerdir. Dani S. vd. [4] yaptıkları çalışmada, seslerdeki duyguyu tespit etmeyi amaçlamışlardır. Çalışmada, K-En Yakın Komşu ve Karar Ağaçları makine öğrenim tekniklerini, Toronto Duygusal Konuşma Seti (TESS) adlı veri seti üzerinde 7 duygu için uygulamışlardır. Sonuç kısmında ise K-En Yakın Komşu tekniğinin %98, Karar Ağaçlarının %92 ve Ekstra-Ağaç sınıflandırıcının %99 doğruluk oranı sağladığını belirtmişlerdir. Ses verisi kullanılarak duygusal durum tespitini hedefleyen Doyranlı Tayşi F. [5], Türkçe 'ye özgü bir sistem geliştirmeyi amaçlamıştır. Çalışmada kızgınlık, üzüntü, heyecan, mutluluk, çaresizlik ve nötr duygularına ait 2194 adet kayıttan oluşan bir veri seti oluşturulmuştur ve çeşitli sınıflandırma algoritmaları kullanılmıştır. Çalışmanın sonuç kısmında veri setleri üzerinde yapılan deneyler sonucunda Rastgele Orman algoritmasının %78,31 doğruluk oranı elde ettiği ve daha başarılı olduğu belirtilmiştir. Korkmaz O.E. [6], ses sinyalinden duygu tespiti yapmayı amaçlamıştır. Çalışmada, EmoDB ve çeşitli veri setleri kullanılmıştır ve Destek Vektör Makinesi ve K-En Yakın Komşu sınıflandırıcılarıyla çapraz doğrulama yapılarak sonuçlar elde edilmiştir. Sonuç kısmında ise Destek Vektör Makinesi algoritmasıyla %98,7 oranında başarı elde edildiğini ve Temel Bileşenler Analizi kullanılarak işlem zamanında ve performansta iyi sonuçlar alındığını belirtmiştir. İnsan sesi verilerindeki olumlu ve olumsuz duyguları tanımlamayı amaçlayan Kao Y. vd. [7] derin öğrenme tekniklerini kullanmıştır. Çalışmada önerdikleri modelin testi için beş duygu veri seti kullanılmıştır. Sonuç kısmında SAVEE veri seti 0.67, TESS veri seti 1.00, RAVDESS veri seti 0.85, IEMO veri seti 0.69, CREMA-D veri seti 0.80 ortalama doğruluk oranı elde etmiştir. Önerilen çoklu model yapısının konuşma duyguları için sınıflandırma doğruluğunu artırabileceği belirtilmiştir.

Çalışmamızı diğer çalışmalardan ayıran ve özgün kılan yönleri şunlardır; duygu tespiti yapılmayan bir veri seti kullanılmış, Türkçe duygu veri seti oluşturulmuş, BERT modeli kullanılarak yapılacak olan Türkçe çalışmalara referans oluşturulmuş ve makine öğrenim algoritmalarının Türkçe veriler üzerindeki başarım oranları tespit edilmiştir.

3. Materyal ve Yöntem (Materials and Methods)

Çalışmanın bu bölümünde veri setine, duygu tespiti için kullanılan BERT modeline ve analiz için kullanılan makine öğrenim algoritmalarına yer verilmiştir.

3.1. Veri Seti (Data Set)

Veri seti, Mozilla tarafından başlatılan ses ve konuşma tanıma yazılımları için ücretsiz bir veri tabanı oluşturulmak üzere geliştirilen *Common Voice* platformundan alınmıştır. Bu platformda, insanların nasıl konuştuklarını makinelere öğretmek amacıyla oluşturulan, çeşitli diller için ses kayıtları ve bu ses kayıtlarına ait metin dokümanları bulunmaktadır. Türkçe ses kayıtlarını içeren 592 MB boyutundaki dosya yaklaşık 20.760 ses verisini içermektedir. Ses kayıtlarının içeriklerini ise; haber videolarından elde edilen sesler ve platform üzerinden yapılan kayıtlar oluşturmaktadır [8].

Birçoğu tekrar eden kayıtlarından oluşan bu ses kayıtlarının tüm metin dokümanları birleştirilmiştir. Elde edilen metin dokümanları üzerinde eleme işlemi gerçekleştirilmiştir. Bir anlam ifade etmeyen, tekrar eden ve sadece sayı belirten veriler ayrıştırıldıktan sonra 5001 adet metin verisi elde edilmiştir. Bu veri seti, üzerinde anlamsız verilerin

ayrıştırılmasından sonra çalışmanın içeriğini oluşturan duygu tespiti ve analiz işlemlerine hazır hale getirilmiştir. Çalışmanın uygulama kısımlarında veri setinin bir kısmı, test için kullanılmıştır. Veri seti üzerinde yapılan oranlama denemeleri sonucunda, test veri seti oranı %15 olarak belirlenmiştir.

3.2. Transformatörlerden Çift Yönlü Kodlayıcı Temsilleri (BERT-Bidirectional Encoder Representations from Transformers)

BERT (Bidirectional Encoder Representations from Transformers-Transformatörlerden Çift Yönlü Kodlayıcı Temsilleri), çok çeşitli doğal dil işleme (NLP) görevleri hakkında en gelişmiş sonuçları elde eden yeni bir eğitim öncesi dil temsili yöntemidir. 2018 yılında Google'da Jacob Devlin vd. tarafından geliştirilen bu model yapısı cümleyi hem sağdan sola hem de soldan sağa taraftan değerlendirmektedir. Böylelikle kelimelerin birbirleriyle olan ilişkilerini daha iyi ortaya koymaktadır. Modelin yapısında, BookCorpus ve Wikipedia veri setleri bulunmaktadır. BERT model yapısı Şekil 1'de verilmiştir [9].

Model, Masked Language Modeling (MLM) ve Next Sentence Prediction (NSP) yöntemleri ile eğitilmektedir. MLM tekniğinde, maskelenen kelime, açık şekilde beslenen kelimelerden yararlanılarak tahmin edilmeye çalışılır. MLM yönteminde daha çok kelimeler arasında ilişkiler üzerinde durulmaktadır. Bir diğer yöntem olan NSP'de ise cümleler arasındaki ilişki üzerinde durulmaktadır. NSP ile model eğitim aşamasındayken iki cümle yapısı arasındaki ilişkiye bakılarak ikinci cümlenin ilkinin devamı olup olmadığına bakılır ve tahmin yapılmaya çalışılır. Eğitim aşamasında gerçekleştirilen optimizasyon ile bu iki yöntem kullanılırken ortaya çıkan kaybın minimuma indirilmesi amaçlanmıştır [9]. BERT model yapısı verilen metin dokümanı hem sağdan hem soldan incelemesi, yapısında bulunan yöntemlerle eğitim aşamasında daha iyi bir öğrenme sağlamaktadır.

3.3. Naive Bayes

Naive Bayes (NB) algoritması, Bayes teoremine dayalı bir öğrenme algoritmasıdır. Tembel (lazy) yapıya sahip olmasına rağmen düzensiz yapıdaki veri setlerinde de çalışabilmektedir [10]. Naive Bayes sınıflandırma algoritması, tüm koşullu olasılıkların çarpımıdır [11]. Bu sınıflandırma algoritmasıyla verilen eğitim ve test verilerinin doğru bir şekilde sınıflandırılması yapılabilmektedir. Naive Bayes algoritmasının kullanım alanları ise şunlardır; spam filtreleme, duyarlılık analizi, çok sınıflı tahmin, öneri sistemleri ve benzeri alanlardır [11].

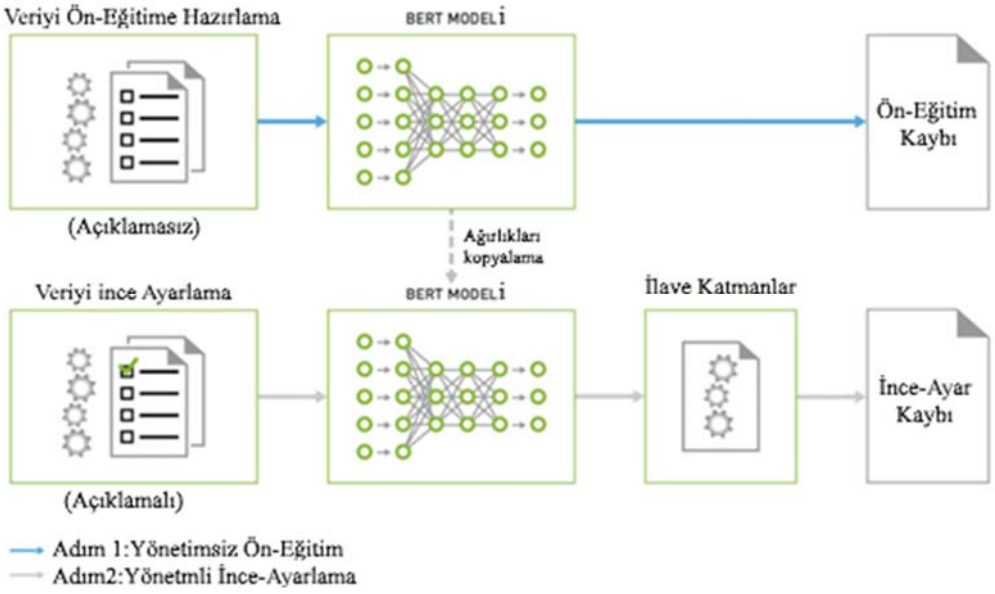
3.4. Rastgele Orman (Random Forest)

Rastgele Orman (RO) sınıflandırma algoritması, birden fazla karar ağacının avantajlarını kullanarak daha uyumlu ve daha iyi sonuçlar üreten modeller ortaya koyarak sınıflandırma işlemini en iyi şekilde yapmaya çalışan bir sınıflandırma algoritmasıdır. Rastgele Orman algoritması çoğu zaman büyük bir sonuç üreten, esnek, kullanımı kolay bir makine öğrenim algoritmasıdır. Rastgele Orman algoritmasının en önemli avantajı ise hem sınıflandırma hem de regresyon işlemlerinde kullanılabilen bir algoritma olmasıdır [12]. Bu algoritma aynı zamanda sınıflandırıcı kategorik değerler için değerlendirilebilmektedir. Rastgele Orman algoritmasıyla ayrıca over-fitting (aşırı uyum) problemlerini de gidermektedir.

3.5. Destek Vektör Makinesi (Support Vector Machine)

Destek Vektör Makinesi (DVM), 1995 yılında Vapnik tarafından istatiksel öğrenme teorisi ve VC-boyut teorisinden türetilmiştir. DVM, ikili sınıflandırma problemini çözmek için geliştirilmiştir. Destek Vektör Makinesi, karmaşık verilerin yüksek doğrulukla işleme özelliği olarak da tanımlanmaktadır. DVM'deki ana amaç ise eğitilmiş veri örneklerini önceden tanımlanmış sayıda sınıfa ayıran, çekirdek fonksiyonu (kernel) denilen lineer olan hiper düzlem (bazen lineer olmayan sınıf ayracı da kullanılır) bulmayı amaçlamaktadır [13, 14].

Bu sınıflandırma algoritmasının avantajları olarak şunlar söylenebilir; yüksek boyutlu uzayda diğer algoritmalara göre daha etkinlerdir, karar fonksiyonu için farklı çekirdek fonksiyon yapıları kullanmaktadırlar ve karar noktalarında kullandıkları eğitim fonksiyonları sayesinde belleğin daha verimli kullanılmasına olanak sağlamaktadırlar [15]. Algoritma kendi içinde de doğrusal ve doğrusal olmayan olarak ikiye ayrılmaktadır. Doğrusal olmayan yöntemde çekirdek fonksiyonları yöntemi kullanılmaktadır. Destek Vektör Makinesi algoritması, metin ve görüntü sınıflandırmada, biyolojik



Şekil 1. BERT Model Yapısı (BERT Model Structure)

bilim alanlarında ve elle yazılmış karakterlerin tanınması gibi alanlarda kullanılmaktadır.

3.6. Karar Ağacı (Decision Tree)

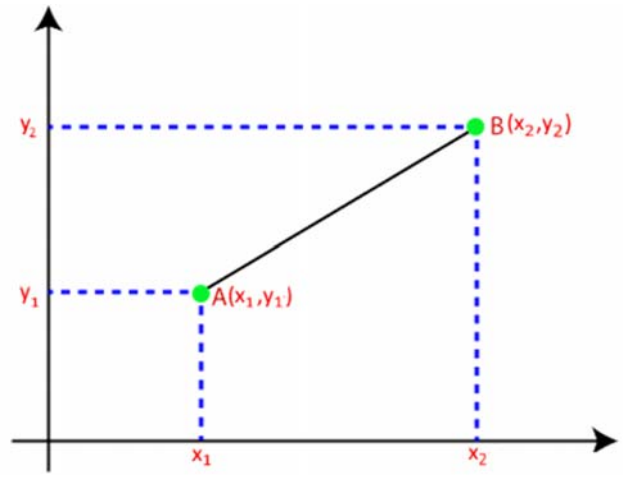
Karar Ağacı (KA), özellik, hedef ve sınıflamaya göre karar düğümlerinden ve yaprak düğümlerinden oluşan ağaç yapısı şeklinde bir model yapısı oluşturan gözetimli sınıflandırma algoritmasıdır [16]. Çeşitli bilim alanlarına uyarlanabilmektedir. Yapısında bulunan en üst düğüme root (kök) ve diğer düğüm yapılarına da leaf (yaprak) adı verilmektedir. Karar Ağacı algoritmaları veri setini küçük parçalara bölerek geliştirilmektedir. Böylelikle algoritma yapısında büyük kayıpların önüne geçilerek daha küçük kayıpların olması sağlanmıştır. Karar ağaçlarının avantajları ise şöyledir; hem kategorik hem de nümerik verileri işleyebilirler, birden fazla çıktısı olan problemler için de çözüm üretebilirler, yorumlanması ve anlaşılması kolaydır ve kullanılan ağaç yapıları görsel halinde sunulabilir [17].

3.7. K-En Yakın Komşu (K- Nearest Neighbour)

K-En Yakın Komşu (KEYK) algoritması, eklenecek olan yeni verinin mevcut veri kümelerine uzaklığını hesaplayarak, k sayıda yakın komşuluğa bakarak sınıflandırma işlemini gerçekleştirmektedir. Uzaklık hesaplamaları için, Euclidean, Minkowski ve Manhattan uzaklık hesaplamalarını kullanmaktadır. Algoritma yapısı eski, gürültülü eğitim verilerine karşı dirençli olduğundan günümüzde halen kullanılmakta olan popüler bir sınıflandırma algoritmasıdır. Algoritma işleyişi ise şu şekildedir; ilk olarak [1, 5] arasında keyfi K parametresi belirlenir, uzaklık hesaplamaları gerçekleştirilir, komşular bulunur ve veri etiketlenmesiyle işleyiş tamamlanır.

Algoritmanın hesaplamalarında kullanılan Öklid (Euclidean) mesafe fonksiyonuna göre (x_1, y_1) ve (x_2, y_2) noktaları arasındaki mesafe Eş. 1'e göre hesaplanır. Bu formülün iki boyutlu grafiği Şekil 2'de verilmiştir.

$$dist((x_2, y_2),(x_1, y_1)) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$



Şekil 2. Öklid Mesafe Fonksiyonu (Euclidean Distance Function)

KEYK algoritmasında, bir K değeri için, algoritma veri noktasının K'ya en yakın komşularını bulacak ve K'nın komşularını en fazla veriye sahip noktalarına atayacaktır. Matematiksel ifadesi Eş. 2'de verilmiştir.

$$P(y = j \mid X = x) = \frac{1}{K} \sum_{i \in A} I(y^{(i)} = j) \quad (2)$$

KEYK algoritması, hem regresyon hem sınıflandırma işlemlerinin yanı sıra birçok alanda sınıflandırma işlemleri içinde kullanılmaktadır [18].

4. Uygulamalar ve Sonuçlar (Applications and Results)

Çalışmamızda ise Türkçe Duygu Tespiti gerçekleştirmek için Türkçe Duygu Tespiti için oluşturulan BERT modelinden yararlanılmıştır. Model yapısı içinde cümleler negatif, pozitif ve nötr şeklinde değerlendirilmektedir. Çalışmada ilk olarak gerekli kütüphanelerin ve veri setinin yüklenmesi işlemleri gerçekleştirilmiştir. Bu işlem adımının sonrasında cümle yapıları tokenlerine ayrıldıktan sonra modele aktarılır ve burada cümlelerin negatiflik, pozitiflik veya nötrlük skorları ve etiketleri belirlenir. Çalışmanın kod kısmında pipeline nesnesi aracılığıyla veriler, peş peşe seri biçimde tanımlanmış birtakım işlemlerden geçer. Bu işlemler sırasıyla verilerin tokenlerine ayrılması ve sonrasında yüklenen BERT modeline aktarılmasını ifade eder.

Veri setinde bulunan cümlelerin karşısında duygu etiketleri (negatif-pozitif-nötr) ve skorları yazdırılmıştır. Gerçekleştirilen duygu tespiti sonrasında 2430 nötr, 1989 negatif ve 582 pozitif duygu etiketine sahip veri elde edilmiştir. Veri seti içinde oranlamaya bakıldığında %48,7 oranında nötr, %39,7 oranında negatif ve %11,6 oranında pozitif cümle bulunmaktadır. Çıktı sonuçlarının bir kısmını gösteren görsel Şekil 3'de verilmiştir.

Duygu etiketleri belirlenen verilerden oluşan veri seti üzerinde makine öğrenim algoritmalarıyla analizler gerçekleştirilmiştir. Bu analizler, Python programlama dili aracılığıyla hem CountVectorizer hem de TF-IDF vektörizasyon yöntemiyle gerçekleştirilmiştir. Her iki yöntemde, çalışmada kullanılan tüm makine öğrenim algoritmalarıyla uygulanmıştır. Veri seti, %85 eğitim verisi ve %15 test verisi olarak ayrılmıştır. Analizler test veri seti üzerinde gerçekleştirilmiştir. Test veri setinde 307 negatif, 75 pozitif ve 369 nötr olmak üzere 751 veri bulunmaktadır. Makine öğrenim algoritmalarıyla yapılan analiz sonuçları dört tablo şeklinde aşağıda verilmiştir. İlk olarak doğruluk oranlarını sonuçları, Tablo 1'de verilmiştir.

Doğruluk oranında elde edilen sonuçlar göz önüne alındığında en iyi sonuca NB ve DVM makine öğrenim algoritmalarında ulaşılmıştır. Doğruluk oranı 0,67 (yüzde 67)'dir. Naive Bayes'de, CountVectorizer yöntemiyle, DVM'de ise TF-IDF yöntemiyle vektörizasyon işlemine tabii tutulan verilerle elde edilmiştir.

Duyarlılık değeri için ilk değerlendirme negatif veriler üzerinde ve CountVectorizer ile elde edilen sonuçlar için yapılmıştır. %67 oranla en iyi duyarlılık değeri oranını RO algoritması elde etmiştir. Negatif veriler üzerinde TF-IDF ile elde edilen sonuçlara bakıldığında en iyi duyarlılık oranı %67 ile DVM algoritması elde etmiştir.

Pozitif veriler için ilk değerlendirme CountVectorizer ile elde edilen sonuçlar için yapılmıştır. %56 oranla en iyi duyarlılık oranını KEYK algoritması elde etmiştir. Pozitif veriler üzerinde TF-IDF ile elde edilen sonuçlara bakıldığında %45 oranla en iyi duyarlılık oranını RO algoritması elde etmiştir.

Nötr veriler için ilk değerlendirme CountVectorizer ile elde edilen sonuçlar için yapılmıştır. %71 oranla en iyi duyarlılık oranını NB algoritması elde etmiştir. Nötr veriler üzerinde TF-IDF ile elde edilen sonuçlara bakıldığında %71 oranla en iyi duyarlılık oranını DVM algoritması elde etmiştir.

Anma değeri için ilk değerlendirme negatif veriler üzerinde ve CountVectorizer ile elde edilen sonuçlar için yapılmıştır. %69 oranla en iyi anma değeri oranını NB algoritması elde etmiştir. Negatif

	Sentence	Sentiment
0	Hayatta küçük şeyleri kovalıyor ve yine küçük ...	-1
1	Seydiu şimdi iki mevkiyi de kaybetti.	-1
2	Fakat öte yandan, hayatta gelişim sağlayabilme...	0
3	Seferler Haziran ayında başlıyor.	0
4	Siyasette temiz insanlara ihtiyacımız var.	1
...	...	...
4996	Kamuoyu ikiye bölünmüş durumda.	-1
4997	Ancak Hiseni suçlamaktan da geri kalmadı.	-1
4998	Kosova başkentindeki yolcu sayısı arttı.	0
4999	Bu tamamen Kosova yönetimine bağlı.	-1
5000	Bu olay daha büyük yankılara yol açacak mı?	-1

Şekil 3. BERT Modeli ile Yapılan Duygu Tespiti Sonuçları (Emotion Detection Results with BERT Model)

veriler üzerinde TF-IDF ile elde edilen sonuçlara bakıldığında en iyi anma değeri oranı %66 oranla DVM algoritması elde etmiştir.

Tablo 1. Makine Öğrenim Algoritmarıyla Elde Edilen Doğruluk Oranı Sonuçları
(Accuracy Results Obtained by ML Algorithms)

Algoritmalar	Doğruluk	
	CountVec (%)	TF-IDF (%)
NB	67	65
RO	64	64
DVM	66	67
KA	61	58
KEYK	54	58

Tablo 2. Makine Öğrenim Algoritmarıyla Elde Edilen Duyarlılık Değerleri Sonuçları
(Sensitivity Values Results Obtained by ML Algorithms)

Algoritmalar	Veri Duygu Durumu	Duyarlılık Değeri	
		CountVec	TF-IDF
NB	Pozitif	0,44	0,00
	Negatif	0,64	0,64
	Nötr	0,71	0,66
RO	Pozitif	0,40	0,45
	Negatif	0,67	0,65
	Nötr	0,64	0,65
DVM	Pozitif	0,39	0,40
	Negatif	0,66	0,67
	Nötr	0,70	0,71
KA	Pozitif	0,35	0,35
	Negatif	0,59	0,56
	Nötr	0,64	0,63
KEYK	Pozitif	0,56	0,33
	Negatif	0,50	0,63
	Nötr	0,58	0,57

Pozitif veriler için ilk değerlendirme CountVectorizer ile elde edilen sonuçlar için yapılmıştır. %56 oranla en iyi anma değeri oranını KEYK algoritması elde etmiştir. Pozitif veriler üzerinde TF-IDF ile elde edilen sonuçlara bakıldığında %29 oranla en iyi anma değeri oranını DVM ve KA algoritmaları elde etmiştir.

Nötr veriler için ilk değerlendirme CountVectorizer ile elde edilen sonuçlar için yapılmıştır. %85 oranla en iyi anma değeri oranını RO algoritması elde etmiştir. Nötr veriler üzerinde TF-IDF ile elde edilen sonuçlara bakıldığında %86 oranla en iyi anma değeri oranını KEYK algoritması elde etmiştir.

F1-skoru için ilk değerlendirme negatif veriler üzerinde ve CountVectorizer ile elde edilen sonuçlar için yapılmıştır. %67 oranla NB algoritması elde etmiştir. Negatif veriler üzerinde TF-IDF ile elde edilen sonuçlara bakıldığında en iyi F1-skoru oranı %67 ile DVM algoritması elde etmiştir.

Tablo 3. Makine Öğrenim Algoritmarıyla Elde Edilen Anma Değeri Sonuçları
(Recall Results Obtained by ML Algorithms)

Algoritmalar	Veri Duygu Durumu	Anma Değeri	
		CountVec	TF-IDF
NB	Pozitif	0,11	0,00
	Negatif	0,69	0,64
	Nötr	0,77	0,80
RO	Pozitif	0,19	0,20
	Negatif	0,50	0,54
	Nötr	0,85	0,82
DVM	Pozitif	0,31	0,29
	Negatif	0,66	0,66
	Nötr	0,73	0,75
KA	Pozitif	0,21	0,29
	Negatif	0,54	0,50
	Nötr	0,74	0,71
KEYK	Pozitif	0,07	0,01
	Negatif	0,54	0,39
	Nötr	0,64	0,86

Tablo 4. Makine Öğrenim Algoritmarıyla Elde Edilen F1-Skoru Sonuçları
(F1-Score Results Obtained by ML Algorithm)

Algoritmalar	Veri Duygu Durumu	F1-Skor	
		CountVec	TF-IDF
NB	Pozitif	0,17	0,00
	Negatif	0,67	0,64
	Nötr	0,74	0,73
RO	Pozitif	0,25	0,28
	Negatif	0,57	0,59
	Nötr	0,73	0,73
DVM	Pozitif	0,34	0,34
	Negatif	0,66	0,67
	Nötr	0,71	0,73
KA	Pozitif	0,26	0,32
	Negatif	0,56	0,53
	Nötr	0,69	0,67
KEYK	Pozitif	0,12	0,03
	Negatif	0,52	0,48
	Nötr	0,61	0,68

Tablo 5. Analizler Sonucunda En İyi Oranlara Ulaşan Algoritmalar (Algorithms Reaching the Best Rates as a result of Analysis)

Veri Duygu Durumu	Duyarlılık Değeri (TF-IDF, CountVectorizer)		Anma Değeri (TF-IDF, CountVectorizer)		F1-Skor (TF-IDF, CountVectorizer)		Doğruluk (TF-IDF, CountVectorizer)	
Negatif	RO	DVM	NB	DVM	NB	DVM		
Pozitif	KEYK	RO	Keyk	DVM, KA	DVM	DVM	NB	DVM
Nötr	NB	DVM	RO	Keyk	NB	DVM, RO, NB		

Pozitif veriler için ilk değerlendirme CountVectorizer ile elde edilen sonuçlar için yapılmıştır. %34 oranla en iyi F1-skorunu DVM algoritması elde etmiştir. Pozitif veriler üzerinde TF-IDF ile elde edilen sonuçlara bakıldığında %34 oranla en iyi F1-skoru oranını DVM algoritması elde etmiştir.

Nötr veriler için ilk değerlendirme CountVectorizer ile elde edilen sonuçlar için yapılmıştır. %74 oranla en iyi F1-skoru oranını NB algoritması elde etmiştir. Nötr veriler üzerinde TF-IDF ile elde edilen sonuçlara bakıldığında %73 oranla en iyi F1-skoru oranını DVM, RO ve NB algoritmaları elde etmiştir.

RapidMiner platformuyla, aynı veriler üzerinde analizler gerçekleştirilmiştir. Vektörizasyon yöntemi olarak TF-IDF yöntemi kullanılmıştır. Analizler test veri seti üzerinde gerçekleştirilmiştir. Her makine öğretim algoritması için operatörlerden oluşan model yapısı oluşturulmuştur. Bu model yapısında her seferinde sadece kullanılan makine öğrenim algoritması değişmiştir.

BERT modeliyle yapılan duygu tespiti sonucunda negatif, pozitif ve nötr veriler / cümleler elde edilmiştir. Bu verilerin, belirlenen makine öğrenim algoritmaları ile RapidMiner ile analizi gerçekleştirilmiştir. Veri seti %85'e - %15 olarak ayrılmıştır. Test veri setinde 298 negatif, 87 pozitif ve 364 nötr olmak üzere 749 veri vardır.

Oluşturulan model yapıları sonrasında çıkan sonuçlar, aşağıdaki tablolarda verilmiştir. İlk olarak Tablo 6'da doğruluk oranları verilmiştir.

Tablo 6. Makine Öğrenim Algoritmarıyla Elde Edilen Doğruluk Oranı Sonuçları
(Accuracy Results Obtained by ML Algorithms)

Algoritmalar	Doğruluk (%)
NB	%60,61
RO	%48,60
DVM	%48,60
KA	%49,27
KEYK	%58,34

Doğruluk oranında elde edilen sonuçlar göz önüne alındığında en iyi sonuca NB algoritmasıyla ulaşılmıştır. Doğruluk oranı %60,61'dir.

Duyarlılık değeri için ilk değerlendirme negatif veriler üzerinde yapılmıştır. %64,56 oranla en iyi duyarlılık değeri oranını KEYK algoritması elde etmiştir. Pozitif veriler için %83,33 oranla en iyi duyarlılık oranını KA algoritması elde etmiştir. Nötr veriler için %66,58 oranla en iyi duyarlılık oranını NB algoritması elde etmiştir.

Anma değeri için ilk değerlendirme negatif veriler üzerinde yapılmıştır. %61,07 oranla en iyi anma değeri oranını NB algoritması elde etmiştir. Pozitif veriler için %24,14 oranla en iyi anma değeri oranını NB algoritması elde etmiştir. Nötr veriler için %100 oranla en iyi anma değeri oranını RO, DVM ve KA algoritmaları elde etmiştir.

Tablo 7. Makine Öğrenim Algoritmarıyla Elde Edilen Duyarlılık Değeri ve Anma Değeri Sonuçları
(Sensitivity Value and Rating Results Obtained by ML Algorithms)

Algoritmalar	Veri Duygu Durumu	Duyarlılık Değeri (%)	Anma Değeri (%)
NB	Pozitif	31,34	24,14
	Negatif	59,67	61,07
	Nötr	66,58	68,96
RO	Pozitif	0	0
	Negatif	0	0
	Nötr	48,60	100
DVM	Pozitif	0	0
	Negatif	0	0
	Nötr	48,60	100
KA	Pozitif	83,33	5,75
	Negatif	0	0
	Nötr	48,99	100
KEYK	Pozitif	43,59	19,54
	Negatif	64,56	34,23
	Nötr	57,61	87,36

Tablo 8. Analizler Sonucunda En İyi Oranlara Ulaşan Algoritmalar
(Algorithms Reaching the Best Rates as a result of Analysis)

Veri Duygu Durumu	Duyarlılık Değeri	Anma Değeri	Doğruluk
Negatif	KEYK	NB	
Pozitif	KA	NB	NB
Nötr	NB	RO, DVM, KA	

5. Değerlendirmeler ve Öneriler (Discussions and Conclusions)

Veriler ve verilerin bir araya geldiği veri kümeleri, son yıllarda araştırma ve analizlerin ilgi odağı haline gelmiştir. Geliştirilen veri analiz platform ve teknikleri de bu durumu tetikleyen etkenlerden olmuştur. Türkçe ses kayıtlarını içeren veri seti üzerinde duygu tespiti ve makine öğrenim algoritmalarıyla analizler gerçekleştirilmiştir. Çalışma sonucunda Türkçe bir duygu veri seti elde edilmiştir.

Python programlama dili ile yapılan analizde; CountVectorizer yöntemiyle elde edilen sonuçlarda Naive Bayes algoritmasının daha çok ön plana çıktığı görülmektedir. TF-IDF yönteminde ise Destek Vektör Makinesi algoritması ön plana çıkmıştır. Vektörizasyon yöntemleri içinde CountVectorizer'da Naive Bayes'in, TF-IDF'de Destek Vektör Makinesi algoritmalarının verimli sonuçlar ürettiği ifade edilebilir. Türkçe verilerle yapılan analizlerde bu iki model yapısının kullanımıyla daha etkin sonuçlar alınabileceği söylenebilir.

RapidMiner platformuyla elde edilen sonuçlara bakıldığında doğruluk oranında en iyi sonucu elde eden Naive-Bayes algoritması özellikle anma değeri metriğinde de etkili olmuştur ve öne çıkmıştır. Bunun yanı sıra Karar Ağacı algoritması da her iki metrikte de iyi sonuçlar elde etmiştir.

Python programlama diliyle yapılan analizde kullanılan Google'ın CoLab ortamı, RapidMiner platformuna göre daha geniş olanaklar sunmuştur. RapidMiner'a göre daha hızlı sonuçlar elde edilmiştir ve birden fazla yöntemin kullanılmasını sağlamıştır. Bu avantaj ise çalışmanın daha iyi analizini sağlamış ve sonuçlar hakkında daha uygun değerlendirmelerin yapılmasına katkı sağlamıştır.

BERT modelinde Türkçe çalışmalar için daha iyi sonuçlar elde edilebilmesi adına BERT modelinin yapısına Türkçe durak kelimeler, Türkçe kelimelerin ve skor yapılarının eklenmesi ve geliştirilmesi için çalışmalar yapılabilir.

Kaynaklar (References)

1. Atalay, M. and Çelik, E., Büyük veri analizinde yapay zekâ ve makine öğrenmesi uygulamalari-artificial intelligence and machine learning applications in big data analysis. Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 9 (22), 155-172, 2017.
2. Demir, A., Atıla, O. and Şengür, A., Deep learning and audio based emotion recognition. In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP) 1-6, IEEE, 2019.
3. Kudiri, K.M., Said, A.M. and Nayan, M.Y., November. Emotion detection using average relative amplitude features through speech. In 2012 IEEE International Conference on Control System, Computing and Engineering, 115-118, IEEE, 2012.
4. Dani S., Mande A.A., Telang S., Shao Z., Emotion Detection Using Audio Data Samples, Int. J. Adv. Res. Comput. Sci., 10 (6), 13–20, Dec. 2019.
5. Tayşi, F.D., Konuşma verisinden duygu durum tespiti (Doctoral dissertation), 2019.
6. Korkmaz, O. E., Ses Sinyalinden Duygu Tanıma, http://acikerisim.ktu.edu.tr/jspui/handle/123456789/531, 2016.
7. Kao, Y.C., Li, C.T., Tai, T.C. and Wang, J.C., 2021, December. Emotional speech analysis based on convolutional neural networks, 2021 9th International Conference on Orange Technology (ICOT), 1-4, IEEE, 2021.
8. Common Voice. https://commonvoice.mozilla.org/tr/about, Yayın tarihi 2021. Erişim tarihi 2021.
9. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
10. Evirgen H., Çerkezi M. Prediction and Diagnosis of Diabetic Retinopathy using Data Mining Technique. TOJSAT. 4 (3), 32-37 2014.
11. Solmaz R., Günay M., Alkan A., Fonksiyonel Tiroit Hastalığı Tanısında Naive Bayes Sınıflandırıcının Kullanılması, Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, 11 (3), 915-924 2021.
12. Kirasich, K., Smith, T., and Sadler, B., Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets, SMU Data Science Review 1 (3), 2018.
13. He, L.M., Kong, F.S., Shen, Z., Multiclass SVM based land cover classification with multisource data, Gastroenterology, 6, 3541-3545, 2005.
14. Kulkarni A.D., Lowe B., Random Forest Algorithm for Land Cover Classification International Journal on Recent and Innovation Trends in Computing and Communication Random Forest Algorithm for Land Cover Classification, 2016.
15. Uçar F., Alçin Ö.F., Dandıl B., Ata F., Power Quality Event Classification Using Least Square-Support Vector Machine, International Conference on Natural Science and Engineering (ICNASE'16) March 19-20, 2016.
16. Onan A., Comparative Performance Analysis of Decision Tree Algorithms in the Corporate Bankruptcy Prediction, Bilişim Teknolojileri Dergisi, 8 (1), Ocak 2015.
17. Demirel Ş., Yakut, S.G., Karar Ağacı Algoritmaları ve Çocuk İşçiliği Üzerine Bir Uygulama. Social Sciences Research Journal, 8 (4), 52-65, 2019.
18. Taşcı, E. and Onan, A., K-en yakın komşu algoritması parametrelerinin sınıflandırma performansı üzerine etkisinin incelenmesi, Akademik Bilişim, 1 (1), 4-18, 2016.